caBIG® cancer Biomedical Informatics Grid®

Considerations for an Enterprise-Wide Semantic Infrastructure for Biomedical Information Management

Prepared for the caBIG® Community

by the

Documentation and Training Workspace

In cooperation with the Vocabulary and Common Data Elements Workspace.

and the

Data Sharing and Intellectual Capital Workspace.

June 2010

Introduction

This document[1] presents an overview of elements and considerations that an organization can include in their plans to establish an enterprise-wide semantic infrastructure. Formally, this paper is aimed at professionals who may oversee, plan or control technical aspects of data sharing, data management, decision support systems, databases, and/or data warehousing within an institution. This can include information technology staff, informatics staff, analysts, implementation/deployment staff, researchers, clinicians or administrators. *Informally*, the paper might be interesting to anyone who has ever had the frustrating experience of needing to create a report or answer a question but was then stymied by not knowing what data resources exist at your institution, who's responsible for those resources, and the process to access or query those resources.

The content presented here was distilled from presentations given at different caBIG® meetings and other related materials. It should *not* be considered a comprehensive review but rather a living document that will change as technologies and the community change (see Feedback section).

The Common Information Landscape – Does this Sound Familiar?

Many organizations have similar issues and problems when it comes to understanding the information landscape at their institutions and these often present barriers to information discovery, stifle data sharing and create redundancies and inefficiencies. The complexity of the biomedical clinical and research domain further magnifies these issues. Familiar examples may include:

- Disparate mix of clinical & research systems chosen with a best-of-breed focus (i.e. electronic medical record systems, lab systems, tissue banking systems, clinical trials management systems, etc) that are blend of commercial, homegrown, or open source systems, and often don't easily interoperate.
- A lot of data buried and inaccessible within unstructured text (progress notes, path reports, etc).

[1] This document is primarily a distillation of two presentations; the first delivered by Michael Riben, M.D. during the CTMS Face to Face Meeting in Feb 2010 titled: "A Practical Roadmap to Support Clinical and Research Systems at M.D. Anderson". The second delivered by David Fenstermacher, Ph.D. during the Deployment Community Face to Face meeting in March 2010 titled: "Making Use of Ontologies & Semantics". This is a living document expected to evolve and change. For details, refer to Feedback section.
http://gforge.nci.nih.gov/svnroot/ctms-forum/CTMS_wide/CTMS-Leads/Face-2-Face/Feb-2010/Presentations/Practical_Roadmap_for_Enterprise_Metadata_to_Support_Clinical_&_Research_Systems_at_MDACC.ppt
https://cabig-kc.nci.nih.gov/MediaWiki/index.php/Making_Use_of_Ontologies

- Need to meet new MediCare/MediCaid Meaningful Use Objectives (EHR Incentive Programs)
- Need to transition to more a more sophisticated level of Service Oriented Architecture (SOA)
- Sparse use of controlled or standard vocabularies
- Little model-driven development
- Little to no data governance program or enterprise standard for coding of terms
- Need to wean off of HL7 v2.0 messaging and migrate to HL7 v3.
- Resource constraints and shifting priorities
- Lack of executive commitment to change management
- Failure to gain organizational adoption of governance

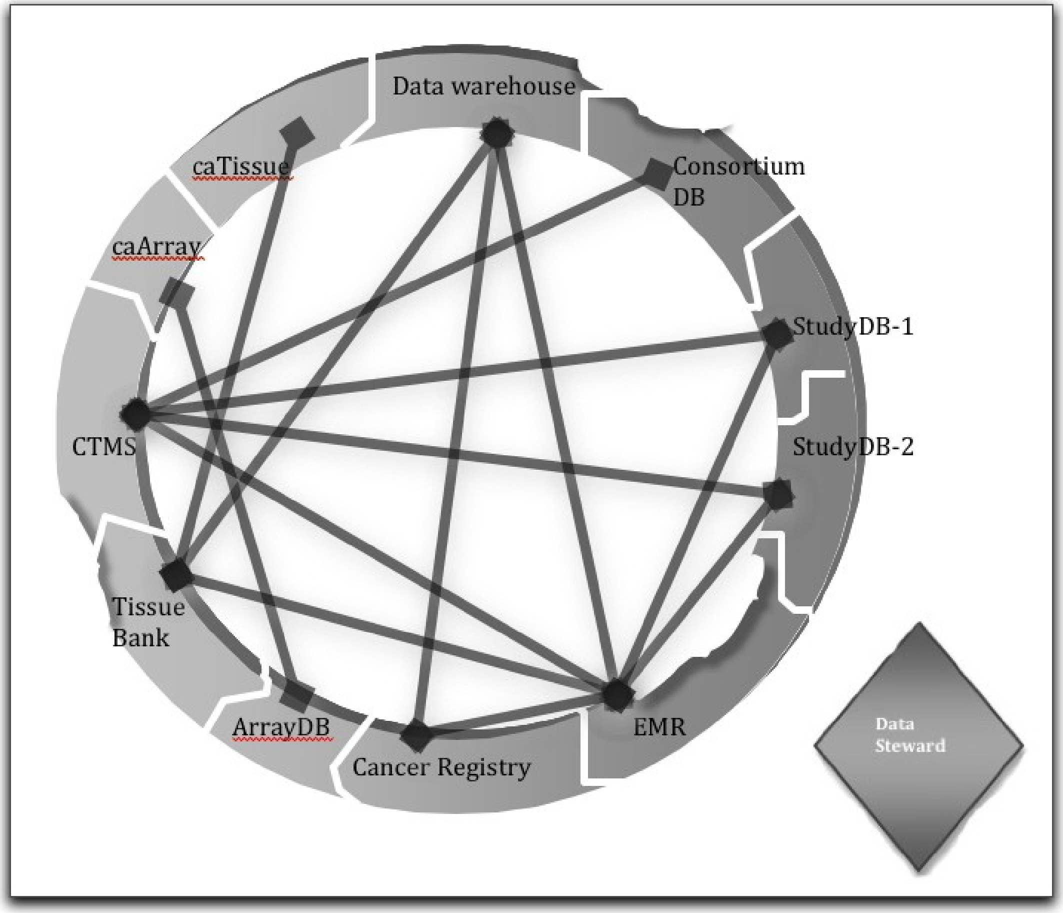


Figure 1: A 'warped bicycle-tire' representation of common information landscape. Spokes between information represent transmissions of data (i.e. point to point web services) and diamonds at the end of spokes represent data stewards or managers.

Why this matters?

While the information landscape described above is common and prevalent, there are a number of compelling reasons to encourage your organization to move towards a more comprehensive, enterprise-wide semantic infrastructure for data management.

- Gain greater efficiencies and minimize re-work by maximizing re-use of terminologies, models and metadata.
- Establishing a semantic infrastructure (e.g. Service-oriented architecture (SOA[2])) could allow greater interoperability between systems, enabling powerful, real-time decision making through, clinical decision support, business intelligence and outcomes reporting.
- Greater awareness and access to data through knowledge services better supports clinical and research missions.
- Aligning with external standards will facilitate greater interoperability and thus data sharing across institutions allowing for greater participation in research consortiums and broader grid initiatives (like caBIG®).

First Steps & Key Elements

Start with the subject matter experts

The subject matter experts are the 'spokes' (Figure 1) to the domain-specific information areas. These individuals exist where the local knowledge lives and have the most vested interest in their data. They are critical to the semantic process. Find out who are the most appropriate individuals within your organization to identify and describe the data needed. Some of their data they identify will need to be harmonized. Engage them to help standardize their data on standard terminologies when feasible. There will likely be some conflict between existing data and standardized data and harmonization will often lead to some amount of data loss though efforts should be taken to minimize this. Former practitioners, cancer registrars, informatics staff are often well suited for this. These individuals can also acts as liaisons between IT and the researchers.

[2] For more information on Service Oriented Architecture (SOA), see the following resources:

- Foster, I. (2005) "Service-Oriented Science" Science 308: 814-817
- The Open Group (2009)"SOA Governance Framework Draft Technical Standard"
- Manes, A.T. (2010) "Understanding SOA Governance" The SOA Magazine (issue XL, June 2010)
- Erl, Thomas. SOA: Principles of Service Design. Boston: Prentice Hall 2008.
- Erl, Thomas. Service-Oriented Architecture: Concepts, Technology and Design. Boston: Prentice Hall 2005.
- Erl, Thomas. SOA: Design Patterns. Boston: Prentice Hall 2009.

Create a local Enterprise Vocabulary Service (EVS)

The local EVS is the best way to continue to serve the subject matter experts, who already have very established terms and ways of naming things while still preserving a mapping to existing standards.

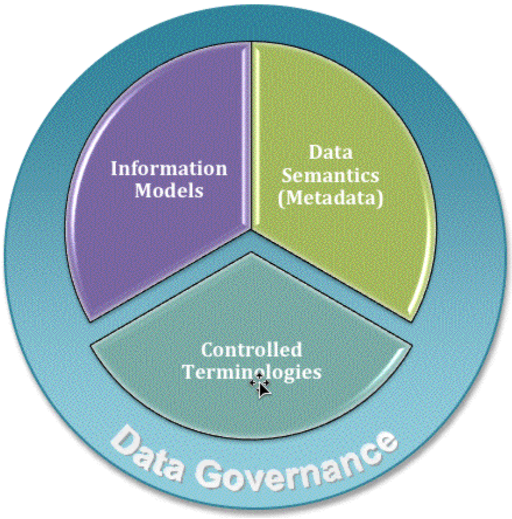
Keep an eye on the BIG picture

When considering an enterprise-wide semantic infrastructure for your institution, be sure to keep a wider view (not just intra-institutional), to also allow for easier interoperability and data sharing with national/international initiatives and consortiums outside your institution (i.e. caBIG®).

Key Elements

The primary components to consider for inclusion in plans for an enterprise semantic infrastructure will be detailed in sections below and include: Controlled Terminologies, Information Models, Data Semantics (Metadata) and Governance. Building a hub comprised of these elements will enable greater discovery, interoperability and data sharing.

Controlled Terminologies: Considerations, Steps & Tools



A terminology can be defined as "A finite, enumerated set of terms intended to convey information unambiguously"[3]. These terms may be associated with one or more concept(s) and can belong in a hierarchy. They may also be mapped or linked to other terminologies (i.e. an interface terminology mapped to a reference terminology). Some examples of standard, controlled terminologies include; International Classification of Disease (ICD), Current Procedural Terminology (CPT), Common Toxicity Criteria (CTC), and Medical Subject Headings (MeSH). Domain specific terminologies and non-standard internal organizational terminologies are common as well.

[3] Source: James J. Cimino, "Principles of Controlled Terminology", presentation: http://courses.mbl.edu/Medical_Informatics/2000.2/Cimino/sld003.htm

Considerations

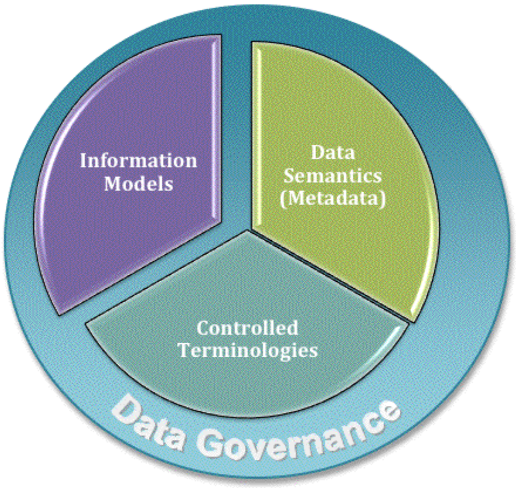
- Consider developing a local enterprise vocabulary service (EVS)
- Encourage re-use of interface and standard terminologies across applications through the use of enterprise services
- Map terminologies to reference terminologies for specific domains such as clinical, research, and administrative
- Consider developing a vocabulary coding system to uniquely identify terms across applications
- Consolidate terminology acquisition, management, development and distribution to avoid redundant purchases
- Integrate terminology and ontology management and development processes in governance plans
- Integrate with modeling and metadata, and support with robust services (i.e. run-time value set services)

Steps & Tools

- Implement a terminology server with robust services - example tool: LexEVS 5.1, HealthLanguage, Apelon
- Load reference and interface terminologies
- Deploy terminology management platform – example tools: Protégé or TopQuadrant , Apelon, healthLanguage, IHTSDO workbench
- Create an "access" methodology – i.e. user access layer

Information Model Management: Considerations, Steps & Tools

An information model can be described as "…a representation of concepts, relationships, constraints, rules, and operations to specify data semantics for a chosen domain of discourse." [4] These representations are not constrained by any specific software implementation. Types of models



[4] Source: Tina Lee (1999). "Information modeling from design to im
of Standards and Technology. http://www.mel.nist.gov/msidlibrar

can include ER (entity relationship) diagrams, or Unified Modeling Language (UML) diagrams. Specific examples of information models used in the biomedical research domain are: HL7 RIM (Health Level 7 Reference Information Model), BRIDG (Biomedical Research Integrated Domain Group).

OWL is a powerful language that can be utilized to assist with information model management. OWL Web Ontology Language is intended to be used when the information contained in documents needs to be processed by applications and can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. OWL has evolved to support the vision for the Semantic Web where machines can perform more useful reasoning tasks on documents and content. Toward that end, OWL has greater facility to express meaning and semantics than XML (Extensible Markup Language), RDF (Resource Description Framework), and RDF-S (RDF-Schema). For example, OWL has more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes.[5]

Considerations

- Consider integrating query and run-time access to application information models through a uniform access layer that is programmatic and user facing
- Develop an enterprise model for clinical and research domains, or commit the standard models that have been developed
- Consider mechanisms to model data buried in unstructured text
- Incorporate automatic metadata generation with semantics
- Include grid access to targeted models

Steps & Tools

- Consider using Natural Language Processing (NLP) to extract valuable information from unstructured text reports – example tools: TRIPS (the Rochester Interactive Planner System) and caTIES
- Establish enterprise standard for model representation for logical, physical metadata - example tool: model in OWL
    - Maintain alignment with NCI and caBIG®

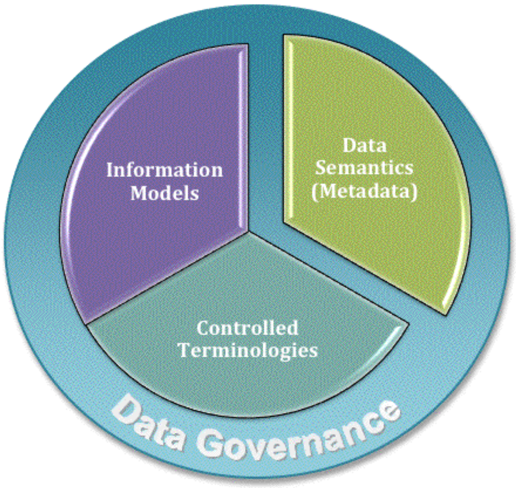
[5] Source: OWL Web Ontology Language Overview. http://www.w3.org/TR/owl-features

- Establish model access methods and promote modeling paradigm for new application development
- Automate model creation/documentation if possible
- Train Developers

Metadata Management: Considerations and Tools

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data.[6]

Developing a framework for handling metadata is one of the key components to a semantic infrastructure. OWL and RDF can be utilized to manage.



Another technique is establishing is a Metadata Registry (MDR). An MDR encompasses data elements, value domains, data element concepts, conceptual domains, and classification schemes.[7] The framework is described in the ISO/IEC 11179 standard. This can assist in providing standardized, single-source semantic definitions for an enterprise.

Considerations

- Consider establishing a semantic Metadata Registry (MDR) that supports both clinical and research applications
- Allow for the creation of services to enable federated information discovery
- Consider a system of federated metadata management for specific domains
- Ensure backward compatibility for ISO/IEC 11179 and HL7 constructs

[6] Understanding Metadata, NISO Press., 2004. http://www.niso.org/publications/press/UnderstandingMetadata.pdf

[7] ISO/IEC 11179, Information Technology –Metadata registries (MDR). http://metadata-stds.org/11179/

Steps & Tools

- Implement an interim centralized metadata registry with goal of federated metadata management – example tools: cgMDR / OpenMDR or Data Foundations Establish metadata access and services
- Consider developing a next generation *MetaModel*
    - Align with NCI's Next Generation Semantic Infrastructure (SAIF/ECCF)
    - Develop or adopt tooling to use OWL as a the modeling construct (HL7 -> OWL based local MetaModel <- ISO/IEC 11179 E2/E3)
- Integrate with your user access layer (UAL) for programmatic access and end-user manual access

Governance: Considerations and Tools

Data governance can be considered the glue or hub that holds the enterprise semantic infrastructure together and as such is a critical piece to establish and can be the most challenging. A model for an enterprise framework for governance should provide a real-time, flexible, simple, system for exercising and enforcement of authority during the entire data life cycle. An effective model includes policies for data management including, quality, roles, and responsibilities for how data collection, data modeling, terminologies and metadata will be managed.

Considerations

- Consider the types of information assets that need governing such as: terminologies, models, metadata, services, etc.
- Consider utilizing a standard for modeling governance workflows
- Integrate, where feasible, already existing compliance/conformance policies and enforcement within departments and groups.
- Consider involving and getting buy-in from senior/executive leadership at your organization (i.e. CIO)

Steps & Tools

- Consider establishing a governance portal - example tools: Sharepoint, Liferay, Collabanet, or a wiki.

- Engage and build consensus for best practices
- Consider establishing workgroups/work streams to focus on key elements of the infrastructure (i.e. Metadata modeling, User Access Layer, Application development, etc.)
- Establish governance touch points and an automated workflow processes for each information resource – example tool: Bonita
- Implement governance for the creation, communication, enforcement, utilization, and adaptation of policies used to direct and control the lifecycle of services in institutional enterprise Service-oriented architecture (SOA).

Enabling Data Sharing and Discovery

When a comprehensive semantic infrastructure is in place, powerful benefits can be realized both internally and externally within the organization. These can be in the form of value-added knowledge services that empower clinicians and researchers with easy and efficient access to data and the ability to further navigate and explore other data services available that they might have otherwise not known about. At an individual level, these benefits might come from a uniform user access layer. From an organizational perspective and beyond, benefits of data sharing can be realized through greater interoperability gained from a service-oriented architecture (SOA).

Considerations

- Try to migrate away from simple point-to-point web services to a service-oriented architecture (SOA).
- Consider establishing a uniform user access layer to the various knowledge services for both internal and external resources.

Steps & Tools

- Establish an SOA with robust Enterprise Service Bus (ESB) and enable discovery, indexing, notifications, mediation etc.
- Establish various knowledge services, such as terminology services, metadata services, federated query services
- Allow services to communicate with
  - caGrid services

    - NCI Enterprise Services (e.g. COPPA)

- Establish a user access layer user (UAL) environment for effective and efficient user interaction with a real-time enterprise knowledge system

- Educate and train the community

- Implement governance for the creation, communication, enforcement, utilization, and adaptation of policies used to direct and control the lifecycle of services in institutional enterprise SOA

- Design and deploy the SOA enterprise framework including the implementation of an enterprise service bus (ESB)

Summary – Tying it Together

The successful implementation of an enterprise-wide semantic infrastructure will likely require a significant level of organizational commitment, time and resources. However, the benefits from this investment should be realized from streamlining the movement of information and facilitating greater data sharing (Figure 2).

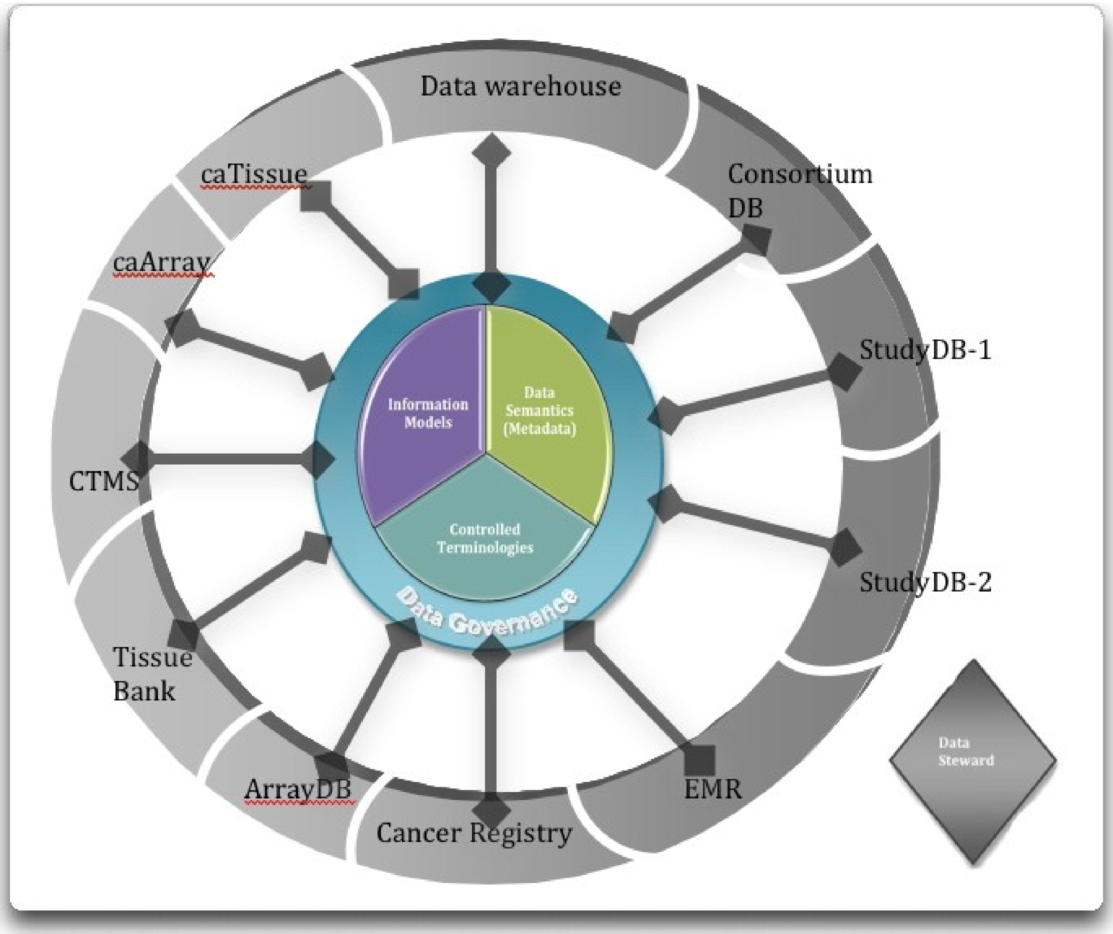


Figure 2: A streamlined 'bicycle tire' representation of an information landscape with a semantic infrastructure in place (hub). The spokes represent data services and the diamonds represent data stewards/managers.

Key components for this infrastructure include information models, controlled terminologies, and data semantics (metadata). Overlaying these components should be a comprehensive, flexible and robust governance model to manage the processes for data management going forward. Together, these components form a more stable, efficient,semantic infrastructure that facilitates and promotes terminology and model re-use and maximizes data discovery and information sharing with positive impacts on clinical care and research.

Feedback

Further questions on the topics covered in this document should be submitted to the appropriate subject forums at the Vocabulary Knowledge Center (VKC). Feedback on this document can be submitted to the General Discussions Forum at that site. https://cabig-kc.nci.nih.gov/Vocab/forums/

Acknowledgements & Credits

Authors

Lara C. Fournier, M.S.
(caBIG® Documentation and Training Workspace)
Informatics Shared Resource
OHSU Knight Cancer Institute
Portland, Oregon

Fred Loney, M.S.
(caBIG® Deployment Lead)
Informatics Shared Resource
OHSU Knight Cancer Institute
Portland, Oregon

Principal Contributors/Editors:

Michael Riben, M.D.
Department of Biomedical Informatics
M.D. Anderson Cancer Center
Houston, Texas

David A. Fenstermacher, Ph.D.
Chair and & Executive Director
Department of Biomedical Informatics
Moffitt Cancer Center
Tampa, Florida